Genomes to Proteomes

Ellen A. Panisko[1], Igor Grigoriev[2], Don S. Daly[1], Bobbie-Jo Webb-Robertson[1], Scott E. Baker[1]

[1]Pacific Northwest National Laboratory, Richland Washington
[2]US DOE Joint Genome Institute, Walnut Creek, California

March 2009

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

DISCLAIMER

Genomes to Proteomes

Ellen A. Panisko[1], Igor Grigoriev[2], Don S. Daly[1], Bobbie-Jo Webb-Robertson[1], Scott E. Baker[1]

[1]Pacific Northwest National Laboratory, Richland Washington
[2]US DOE Joint Genome Institute, Walnut Creek, California

1. Introduction

Biologists are awash with genomic sequence data. In large part, this is due to the rapid acceleration in the generation of DNA sequence that occurred as public and private research institutes raced to sequence the human genome. In parallel with the large human genome effort, mostly smaller genomes of other important model organisms were sequenced. Projects following on these initial efforts have made use of technological advances and the DNA sequencing infrastructure that was built for the human and other organism genome projects. As a result, the genome sequences of many organisms are available in high quality draft form.

While in many ways this is good news, there are limitations to the biological insights that can be gleaned from DNA sequences alone; genome sequences offer only a bird's eye view of the biological processes endemic to an organism or community. Fortunately, the genome sequences now being produced at such a high rate can serve as the foundation for other global experimental platforms such as proteomics. Proteomic methods offer a "snapshot" of the proteins present at a point in time for a given biological sample. Current global proteomics methods combine enzymatic digestion, separations, mass spectrometry and database searching for peptide identification. One key aspect of proteomics is the prediction of peptide sequences from mass spectrometry data. "Global" proteomic analysis uses computational matching of experimental mass spectra with predicted spectra based on databases of gene models that are often generated computationally. Thus, the quality of gene models predicted from a genome sequence is crucial in the generation of high quality peptide identifications. Once peptides are identified they can be assigned to their parent protein. Proteins identified as expressed in a given experiment are most useful when compared to other expressed proteins in a larger biological context or biochemical pathway.

In this chapter we will discuss the automatic annotation and the generation of high quality gene models, the setup and execution of global proteomics experiments that are quantitative and statistically rigorous and finally add biological context to proteomics.

2. Gene modeling

Genome sequencing has evolved dramatically in the last few years. The sequence of the first bacterial genome of *Haemophilus influenzae* was published in 1995 (Fleischmann et al, 1995) shortly followed by the first sequenced eukaryotic genome of *Saccharomyces cerevisiae* (Goffeau et al. 1996). Several large genome sequencing centers were

established around the world for large scale production sequencing and have an average sequencing capacity of 3Gbases per month, or roughly an equivalent of the human genome size. New short-read sequencing technologies promise to make genome sequence affordable for small laboratories and research groups. Anticipated massive amounts of sequence data require adequate efforts and tools for analysis and interpretation of these data. Genome annotation is one of the first steps in analysis of genome sequence and includes finding genes, describing their structures and functions. Approaches used for gene prediction in prokaryotes and eukaryotes are different. Finding genes in prokaryotes is relatively straightforward task because of simple gene structure (uninterrupted open reading frame, or ORFs) and high gene density, with almost the entire DNA used for coding. In contrast, eukaryotic genes have complex exon-intron structures and a significant fraction of eukaryotic genome sequence corresponds to non-coding DNA (for example, gene deserts in human (Taylor, 2005)).

Despite significant efforts in many research groups, unlike in prokaryotes, there are no completely automated methods to predict gene models in eukaryotic genomes. Most of the eukaryotic gene predictors that have been developed and tuned for human or other higher eukaryote genomes, are not applicable to another genome, and show low accuracy even between vertebrate genomes (Burset and Guigo 1996). Eukaryotic gene predictors require training for every organism on a set of known genes from that organism's genome. This information is used to derive genome-specific parameters that then are used to predict genes in whole genome. Several benchmarks were developed to evaluate current gene predictors for human (EGASP, Guigo et al, 2007), fruit fly (GASP, Reese et al, 2000), maize (Yao et al, 2005), and other genomes (e.g., NGASP, www.wormbase.org).

Gene predictors

Eukaryotic gene predictors can be roughly described as *ab initio* (for example, Fgenesh, Salamov & Solovyev, 2000; Augustus, Stanke & Waack, 2003; SNAP, Korf, 2004; GeneMark, Lukashin & Borodovsky, 1998), homology-based (GeneWise, Birney & Durbin, 2000; Fgenesh+, Salamov & Solovyev, 2000), EST-based (GrailEXP, Xu et al., 1997; PASA, Haas et al., 2003), synteny-based (Twinscan, Tenney et al., 2004), and hybrid methods (EuGene, Schiex et al., 2001; Combiner Allen et al., 2004; TWAIN, Majoros et al., 2005)). They differ in balance between *content-based* (distinguishing exons from introns or intergenic regions by, for example, nucleotide composition) and *signal-based* parameters (defining starts and ends of exons and genes) (Solovyev, 2002). The *content* information can come from homology to proteins, ESTs and genome conservation as well as coding potentials derived from a training set of genes. *Signals* while mostly conserved can be refined based on homology gene models and ESTs aligned to genomic sequence. In general, the predicted models will be highly inaccurate if the genome that the gene finding algorithm is applied to is different in gene structure than the genome that the algorithm was trained on (Korf, 2004).

Given a sufficient number of known genes or full-length cDNAs for a particular genome, gene prediction parameters can be computed and used for genome-wide gene prediction. Often for the most of newly sequenced genomes we do not have full-length cDNAs and

some characteristics of gene structure in a given genome can be inferred from ESTs. They can be directly mapped to genome assembly or used in EST-based gene predictors such as PASA (Haas et al., 2003). Reliable homology-based gene models built with GeneWise (Birney & Durbin, 2000) or Fgenesh+ (Salamov & Solovyev, 2000) offer another source of information for training gene predictors. While these predictions lack UTR regions, close protein homologs often retain very similar exon-intron structures. In addition, genomes of closely related organisms can help to recover content and signal information using synteny-based gene prediction methods. These methods were successfully used in human, mouse and rat (SLAM, Dewey et al 2004) *C.elegans* (TwinScan, Wei et al, 2005), *Aspergillus* genomes (TWAIN, Majoros et al. 2005), *Cryptococcus neomorphans* (TwinScan, Tenney et al. 2004), and *Phytophthoras* (Tyler et al., 2006). These methods predict exons with a reasonable quality but suffer from chimerism in genome scale application and, therefore, often used mostly to correct models of orthologous genes.

Annotation pipelines

Since each gene prediction method has its own advantages and drawbacks combining different methods can improve overall quality of gene models. Methods to select entire gene models such as Bayesian framework (Pavlovic et al. 2002), to assemble model fragments into *de novo* models (e.g., EuGene (Schiex et al, 2001), or to combine multiple sources of information such as gene models and ESTs (Allen et al. 2004)) have been proposed. Annotation pipelines employed at the genome sequencing centers usually use several gene predictors. In addition to increasing overall accuracy of annotations, they offer scalable solutions. ENSEMBL pipeline was used for most of vertebrates genomes (Potter et al, 2004). The US DOE Joint Genome Institute (JGI) Annotation Pipeline includes Fgenesh (Salamov & Solovyev, 2000), GeneWise (Birney & Durbin. 2000), and Fgenesh+ (Salamov & Solovyev, 2000) with a number of in-house developments to use ESTs and select a best representative model for every locus among the number of predicted genes. The Broad Institute used similar set of tools for annotation of fungal genomes. The Institute for Genome Research (TIGR) /J. Craig Venter Institute (JCVI) annotation team trains several gene predictors but use a subset of them for final annotations. Genome sequencing consortia use additional gene predictors (Dujon et al. 2004,Braun et al. 2005, Jaillon et al, 2007).

After gene models are predicted, the corresponding predicted proteins are functionally annotated. Functions can be inferred by sequences similarity to other proteins from, for example, UniProt or GenBank as determined by protein sequence alignments using Blast (Altschul et al., 1998). InterProScan (Zdobnov et al., 2001) combines several domain-search methods to predict domains including SignalP and TargetP (Emanuelsson et al. 2007) for more specialized analysis. Comparison with the specialized databases (e.g., KEGG (Kanehisa et al. 2004)) allows one to map the predicted proteins onto metabolic pathways, Gene Ontology (Ashbruner et al, 2000) and KOG (Tatusov et al. 2003) categories provide the user with multiple entry points into the annotation data.

The overall workflow is similar between the different pipelines and includes the following major steps common to all:

1. Repeat masking to exclude transposons from the final set of gene models,
2. Mapping ESTs and homologs as seeds for gene predictors,
3. Gene prediction using several methods,
4. Gene annotation via domain prediction and homology searches.

Additional experimental data available at the time of genome annotation is becoming more often an integral part of validation modules of these pipelines.

Experimental validation and annotation of predicted gene models

Accuracy of predicted genes depends on derived parameters and varies from genome to genome. A significant fraction of predicted genes with no similarity to any protein in GenBank lacks annotation. Experimentally derived data (ESTs, microarrays, proteomics) can not only validate predicted gene structures but also add annotation by describing conditions under which a particular gene or protein was expressed. Predicted gene models can be validated using gene expression data. Evidence for predicted transcripts can be collected from ESTs/cDNAs overlapping with a gene model, microarrays with oligonucleotide probes corresponding to the predicted transcripts, and tiling arrays where probes are evenly distributed throughout the genome sequences. In addition comparative analysis of ESTs from different libraries or microarray probe hybridization level under different conditions provides biological insights and annotation information. These resources are stored in genome databases as well as larger repositories (ArrayExpress, Parkinson et al, 2005; GEO, Barrett et al, 2005). In addition to a wide variety of proteomics biomedical studies, other examples include secreted proteins in fungi that degrade of biomass (Medina et al. 2005, Medina et al. 2004, Vanden Wymelenberg et al. 2005), or have symbiotic relationships with plants (Bestel-Corre et al. 2004; Martin et al, 2008). For example, 10,048 genes were predicted for the genome of *Phanerochaette crysosporium* using 10.6x genome sequence assembly processed with the JGI Annotation Pipeline. Processing mass-spectroscopy data by running against MASCOT db resulted in identification of 4,697 peptide supporting 1,489 genes including 193 peptides supporting splice sites. Genome browser of the JGI Genome Portal illustrates peptide support for predicted gene models (Fig 1) (Zhou et al, in press).

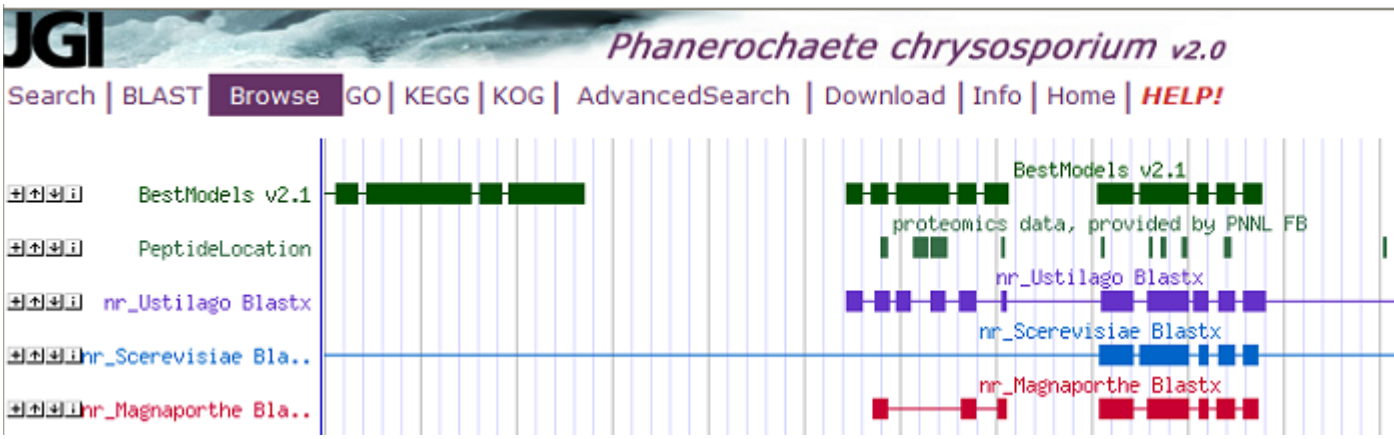

Figure1. Peptides mapped to genome assembly provide experimental support for predicted gene models in *Phanerochaete chrysosporium* genome

Challenging genes that require validation with proteomics
While proteomics is valuable in validating predictions of protein coding genes, an additional value comes from its ability to distinguish between protein coding and non-coding genes, transcripts for both of which can be equally supported by ESTs or microarrays.

Pseudogenes:. Remnants of genes that are no longer transcriptionally active are called *pseudogenes*. Based on their origin they are subdivided into *processed* (emerged trough retrotransposition of processed transcripts back into genomic sequence) and *nonprocessed (*duplicated, not active and therefore mutated genes). Pseudogenes often have features that make them appear to be genes. Sometimes they are expressed based on EST or microarray evidence. Increased rates of mutation can introduce stop codons or frameshifts. The frameshifts can be result of either sequencing error or genomic mutation, especially for non-expressed genes, and possible be resolved with proteomics.

Seleno proteins: Selenocysteine (Sec), a rare amino acid that significantly increases enzymatic activity of a protein, is coded by a nucleotide triplet UAG normally interpreted as a stop codon. In presence of cis-acting mRNA structure, called selenocysteine insertion sequence (SECIS) element, this codon is recognized by seleno-cystein tRNA, which integrates a Sec amino acid into protein sequence. Presence of stop codon in the middle of sequence of a predicted gene/transcript makes it a viable pseudogene candidate, but since some pseudogenes are expressed in form of RNA, only protein expression can support this type of proteins.

Non-coding genes: Often clusters of ESTs suggest missing genes in places where no gene model was predicted. Lack of a gene model indicates lack of significant coding potential, homology in that locus, which could be due to incorrect training or a specific genes. Finding ORF does not necessarily mean coding genes since a long ORF can be found even in non-coding RNAs.

Polycystronic genes: proteomics data can resolve conflict between different ORFs found in the same genes, either in gene in low GC w/o any stops, or in polycystronic genes, where several genes are expressed as the same transcript to be processed before translation.

Caveat: Resolving mass specta requires a database of protein sequences which are derived from predicted gene models. To support predicted gene models we use the same gene models to resolve mass spectra. One option is to this is use all ORFs in 6 frame translation derived directly from genomic sequence. However, exon-intron structure of eukaryotic genes make this difficult. Only peptides that align entirely within a single exon can be resolved this way and peptides aligned across a splice site will be lost.

3. Proteomics Experimental Design

In the majority of proteome studies investigators, are interested in comparing the proteins expressed in a cell or tissue under one condition versus another (i.e. normal vs. diseased).

Originally, proteome studies were conducted using two-dimensional polyacrylamide gel electrophoreses (2D-PAGE, Righetti et al., 2004). In this method, proteins are separated by isoelectric point and then by size. After running the sample, the gel is stained with a protein binding dye. Image analysis is performed to compare each stained gel from the two (or more) conditions. A 'spot" of interest can be excised from the gel and the protein identified using mass spectrometry (Gevaert and Vandekerckhove, 2000). Technology has been developed to allow two samples to be analyzed on a single gel by labeling each sample with a different fluorescent dye (Marouga et al., 2005). However, currently the field of proteomics is dominated by relatively high-throughput mass spectrometry based approaches.

In these methods, high pressure liquid chromatographs (HPLC) are coupled directly to mass spectrometers (MS), as peptides elute from the HPLC column they are converted to the gas phase by electrospray ionization (Cole, 2000; Griffiths et al., 2001) and drawn into the inlet of the MS. Currently, different types of MS and HPLC platforms are used across proteomics focused laboratories. For the basic shotgun approach to proteomics, peptides isolated from cells are digested by a protease (typically trypsin) with defined cleavage sites. The resulting peptides are analyzed by HPLC/MS techniques. For an example of the complexity of the resulting sample, we consider the worm *Caenorhabditis elegans*. The *C. elegans* genome encodes approximately 20,000 ORFs which theoretically can produce close to a million tryptic peptides (Conrads et al., 2000). Since tryptic digest samples are too complex to resolve each individual peptide in time by HPLC, an identical sample injected run through a HPLC/MS system will have a limited number of overlap of identifications (this is often referred to as undersampling, Figure 2).

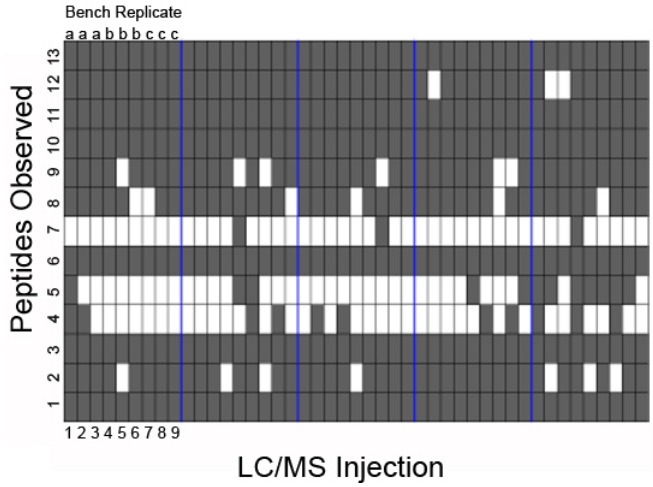

Figure 2. Peptides observed from NADP-dependent isocitrate dehydrogenase in total soluble proteome analyses of Trichoderma reesei. Each column represents an individual LC/MS injection of sample and each row represents a peptide from isocitrate dehydrogenase that has been observed at some time in previous experiments. White indicates the peptide was observed in the sample and grey indicates the peptide was not observed. Blue lines separate five different samples examined. Each sample has three bench top replicates, each of those replicates had three LC/MS technical replicates resulting in nine injections for each of the five samples.

Determining differences in protein expression between samples is less intuitive than with the 2D-PAGE method and there are several methods utilized in current research (reviewed in Bantsacheff et al. (2007) and Nesvizhskii et al. (2007)). Perhaps the most straightforward is called “spectral counting”, which entails counting the number of spectra a peptide (or peptides) from a protein was observed in during a HPLC/MS analysis (Mann and Wilm, 1994). The counts from two different sample types are compared to identify proteins that are differentially expressed.

For model systems with defined growth media, stable isotope labeling strategies are often performed (Ong et al., 2002). In these studies, one cell condition (normal) may be grown in baseline media and the other cell is grown with media where stable heavy isotope labeled amino acids are substituted. Equivalent amounts of cells or protein extracts from the two cells are combined and processed for analysis. The ratio of mass spectral intensities between the heavy and light isotopically labeled peptides is used for relative quantitation. For cells with undefined growth media such as human tissue samples, a similar strategy can be used with affinity labels (ICAT, Gygi et al., iTRAQ, Ross et al., 2004). The affinity tag is produced in two versions, heavy and light. The protein extract from one condition is treated with the heavy reagent and extract from the other condition of interest with the light reagent. Equivalent amounts of labeled extract are them combined for processing and LC/MS analysis. One advantage of affinity labeling methods is that they can isolate specific peptides (with ICAT only cysteine containing peptides), thereby reducing overall sample complexity.

Software such as MASIC (Monroe, 2006) is used to determine the mass spectral intensities of peptides. The process starts with the parent ion (mass to charge ratio) that was identified for a peptide and extracts the elution profile (extracted ion chromatogram) of that ion from the mass spectra collected for that LC/MS injection. Essentially this is a plot of the parent ion intensity over time. The peak area and maximum peak intensity can be calculated and used for quantitation.

The same method can be applied to non-isotopically labeled samples for relative quantitation. Non-labeled experiments are not limited by the number of heavy isotopically labeled amino acids or affinity tags available. But regardless of method utilized, all experiments benefit from a strong experimental design. Caution must be taken to reduce sample preparation variability and prevent experimental processing from biasing data. Experimental bias can result from preparing all “like” samples together and separate from the remaining conditions of an experiment or having only one replicate per sample.

4. Proteomics Sample Processing

It should not be surprising given the various methods for proteome studies, that there are also a myriad of approaches for sample processing. Often an investigator will focus on isolating a specific type of protein from a sample. For example if an investigator is interested in isolating only phosphorylated proteins, they can choose from an affinity

labeling technique (PhIAT, Goshe et al., 2001) or a chromatographic method (IMAC, Ficarro, 2002). Too numerous to discuss here, sample processing methods are reviewed by Canas et al. (2007) and Bodzon-Kilakowska et al. (2007). Here we outline a basic procedure for total soluble protein proteome sample processing.

After harvesting the cells or tissues, samples are typically stored frozen until all biological replicates can be processed in parallel (or according to your experimental design). Depending on sampling techniques, experimental (bench) replicates can be initated either before or after cell lysis. For example, if one can easily determine cell number, replicate samples can be produced by putting an equal number of cells into separate tubes. Each tube will be processed separately through out the entire method. Cells may be lysed chemically or mechanically depending on the model system employed. Lysis is often done in the presence of a high molaritiy chaotrophic salt such as urea or guanidine to denature proteins as soon as the cell contents are released. Protease inhibitors may also be added to the lysis buffer. Cell debris is then removed from the samples by centrifugation and the supernatant is reserved for further processing.
The cell lysate is subsequently assayed to determine protein concentration usually using the BCA (Smith et al., 1985) method due to its tolerance of high salt. Experimental replicates may also be started at this stage for those systems where cell number is not easily assayed, simply by aliquoting equivalent amounts of protein to separate tubes. Many researchers also denature the protein sample by incubating with TCEP (Tris(2-carboxyethyl) phoshpine) and chemically modify cysteines by incubation with iodoacetamide to prevent disulfide bond formation. Samples are now ready for tryptic digestion.

For effective digestion, samples must be diluted with buffer to reduce the concentration of salt to a level tolerated by the enzyme, and ensure the sample is at the appropriate pH. Trypsin is added to the sample in a ratio of anywhere from 1 part trypsin and 20 to 100 parts sample protein. Trypsin is prepared according to manufacturers instructions added to the sample and incubated at 37° C overnight.

Before processing by HPLC/MS peptides from the sample are separated from salts and concentrated using solid phase extraction. This procedure uses a matrix of silica beads to which are attached chains of hydrocarbon 18 carbons in length (reverse phase). Solid phase extraction cartridges are typically made to attached to a vacuum manifold that allows liquid to be pulled through the column and provide a place to collect the final eluate of peptides. The resin is first activated by adding 2-4 column volumes of methanol. After activation, the column is not allowed to dry. Then water (2-4 column volumes) is added to equilibrate the resin to aqueous conditions. Sample is then added to the column and washed with 6 to 8 column volumes of water or volatile buffer (ammonium bicarbonate). Peptides are eluted with 2 column volumes of organic solvent, often 80 to 100% acetonitrile. The samples are dried using a centrifugal vacuum concentrator and can be stored frozen until HPLC/MS analysis.

It is difficult to describe a “typical” HPLC/MS experiment as techniques to improve the chromatographic separation of peptides and detection by mass spectrometry is a very

active area of research. Usually reverse phase (essentially separation of peptides by hydrophobicity) chromatography is used and eluate from the HPLC column is injected directly into the mass spectrometer in real time. Multidimensional separations can also be performed where the sample is separated in one dimension by strong cation exchange into fractions which are then subjected to reverse phase separation. It is possible to do this "on-line" with a single biphasic column (Washburn et al. 2001), as well as separately.

The columns used are fused silica with inner diameters in 75 to 150 microns which are packed with 3-5 micron reverse phase particles. After the column is equilibrated in aqueous with dilute volatile acid (ensuring peptides will be positively charged) samples can be injected onto the column. The peptide samples isolated by solid phase extraction are resuspended in the same solvent used to equilibrate the column. Peptides will be eluted off the column by adding increasing amounts of organic solvent containing the same acid. Gradient profiles used vary across investigators, one example is shown in Figure 3.

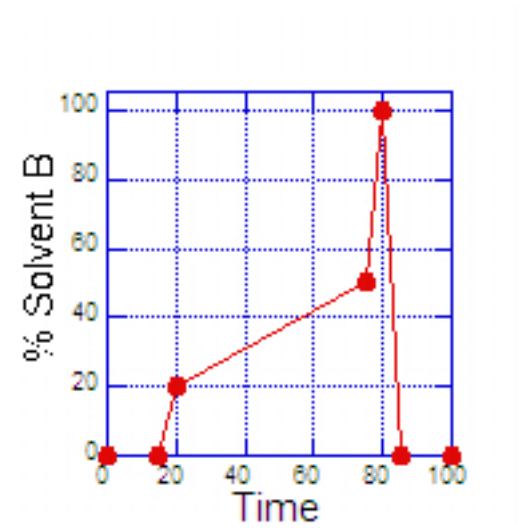

Figure 3. An example of a solvent gradient used for eluting peptides from a HPLC column. For 15 minutes after the sample in injected the column remains at 100% solvent A (0.1% formic acid). Then a linear gradient from 100% to 20% solvent B (90%acetonitrile, 0.1% formic acid) over 5 minutes, followed by another linear gradient from 20% solvent B to 50% solvent B over 55 minutes and finally a linear gradient from 50% solvent B to 95% solvent B over 5 minutes. Before the next sample is injected the column must be reequilibrated in solvent A.

Mass spectrometers are typically run in a data dependent mode, choosing the most intense ions observed in a survey (MS) scan isolating those ions for fragmentation (MS/MS scan) in subsequent scans. Thousands of scans can be collected for a single HPLC/MS injection. Since the manner of peptide fragmentation is predictable, the mass to charge ratio from the survey scan and the ions produced by fragmentation are used for identification. Software programs like Sequest (Eng et al., 1994) and X!tandem (Craig and Beavis, 2003) perform probability based identification by utilizing the data from the

mass spectrometer and comparing it to the expected peptide fragmentation for all tryptic peptides from an *in silico* tryptic digest of all proteins from a defined protein database (i.e., SwissPro or the predicted proteins from an organism whose genome has been sequenced).

Data analysis is an extremely important aspect of proteome studies, which deserves more than the cursory mention that we include here. Often an experiment will involve dozens if not hundreds of HPLC/MS injections, where data files are large. Thousands of peptides are identified in a single analysis and laboratories which specialize in proteomics often have their own data management systems (Kiebel et al., 2006).

5. Statistical Modeling of Proteomics Data

A biological study of peptides by liquid chromatography (LC) coupled with mass spectrometry (MS) produces a large, complex, but sparse data set due to the design of the study, the LC/MS queuing plan for study samples, and the (often incomplete) observation of numerous peptides across the study's sample collection. Consider a study of cells grown with exposure to five different concentrations of pesticide, each condition having multiple samples with replicate LC/MS injections. The realized design has an intricate structure spanning thousands of peptide measurements perforated with missing observations. To ensure valid and objective biological conclusions, a statistical method for an LC/MS-based biological study should formulate a design-complementing queuing plan that complements experimental design so that data suitable for the appropriate statistical modeling is collected. Then a matching statistical analysis can be performed. Statistical modeling is defining, fitting and interpreting a probability model. The simplest statistical algorithm is an exercise in statistical modeling if the intention is to make inferences about problems behind the data. Under each application, a (potentially invalid) probability model implicitly looms. The validity of this exercise depends upon an understanding and application of basic statistical concepts that underpin designing, fitting and interpreting probability models. Numerous internet resources offer quick, outstanding refreshers about important basic statistical concepts. These include Wikipedia (Wikipedia, 2008), NIST SEMATECH e-Handbook of Statistical Methods (NIST, 2008), Electronic Statistics Textbook (StatSoft, Inc., 2008), EBook (UCLA Department of Statistics, 2008) and MathWorld Probability and Statistics (Wolfram, Inc. 2008).

Mixed-effects Modeling

Mixed-effects linear statistical modeling (Pinheiro and Bates, 2000) is an established statistical methodology for the analysis of comparative, screening and time course experiments. A mixed-effects model includes terms for both fixed effects such as researcher-set treatments, and random effects due to subject response, instrument variability or other nuisance factors. The mixed-effects modeling approach uniquely suited to producing a LC/MS sample queuing plan and statistical analysis complementary to the often complex realized design of a biological study. A detailed discussion and example is described in Daly et al. (2008). Here we offer a brief overview.
The basic steps are:

1. Identify the LC/MS nuisance factors,

2. Evaluate the design of the biological study,
3. formulate the LC/MS queuing plan,
4. explore the LC/MS data set,
5. define and fit protein-level mixed effects models,
6. group proteins based on estimates of biological parameters,
7. draw biological conclusions about individual proteins and protein groups,
8. alternatively, draw conclusions about the quality and performance of the LC/MS process.

The mixed effects statistical model is characterized by three important elements. First, the model describes an LC/MS abundance measurement as a multiplicative function of study and processing factors. To facilitate modeling fitting, this multiplicative model is log transformed so that log(peptide abundance) is expressed as an additive model of study and processing factors. The model generates estimates of model goodness-of-fit, treatment and peptide effects, standard errors, and confidence intervals. Pertinent results are then transformed back to the original scale for biological interpretation. Second, the model has two disparate sets of terms-one set represents the biological design while a separate set represents the LC/MS sample processing plan. Third, the relative difference in LC/MS measurability between peptides is represented by a component measurability factor.

A biologically induced difference between two conditions is often inferred from the ratio of a peptide's LC/MS abundance estimates (i.e., a component's relative abundance). The acceptance of forming this ratio to eliminate or significantly minimize the systematic effects of LC/MS processing, coupled with the common assumption that measurement error is relative (i.e., MS measurement error increases with measurement value) suggests that variation in MS abundances may be explained adequately with a multiplicative error probability model. Further, sample effects due to dilution/titration, fractionation, etc., are often multiplicative in nature. Consequently, log-transformed MS abundances may be effectively described by an additive statistical model (i.e., in matrix notation, model terms and coefficients are separable).

Restricted maximum likelihood estimation (REML) is the method used for model fitting. REML was developed and refined to estimate more accurately variance components in random and mixed-effects models (Patterson and Thompson, 1971; Laird and Ware, 1982; Searle et al., 1992). REML correctly tabulates the degrees of freedom for unbalanced data, improving error estimates and inferences. REML is better suited to fitting linear models to the often incomplete LC/MS datasets than other techniques such as ordinary least squares analysis or analysis of variance.

Data Quality Issues

Variance in LC/MS analysis is a significant challenge. Ideally, each protein which is processed would be extracted, digested, purified, separated by LC and observed by MS with equal efficiency. Proteins that are equimolar in a sample, would have comparable MS abundances proportional to their concentration, and in particular, abundances peptides from their parent protein would be but replicate measurements of the parent

protein's abundance. In reality LC/MS processing, some peptides are more easily measured (i.e., identified and quantified) by LC/MS than others (Purvine et al., 2004). Whether caused by peptide digestion efficacy, hydrophobicity, or ionization potential these nuisance factors directly affect the quantification of a component's abundance. This LC/MS peptide measurability effect varies across peptide due to the cumulative, but differential effects of nuisance factors. Relative LC/MS peptide measurability, however, is very reliable across samples measured under similar conditions on the same LC/MS platform. That is unique peptides of a given protein most often display similar MS abundance profiles randomly perturbed by measurement error across a biological study. Differences in the LC/MS peptide measurability can be estimated and removed by mixed-effects modeling to remove this source of variability and allow pooling of data from peptides of the same parent protein (Daly et al, 2008). The mixed effects modeling produces one model for each fitted protein. A single study may result in hundreds to a few thousand acceptable individual protein models.

LC/MS processing introduces many nuisance factors unrelated to the biological factors of greatest interest, such as variability in instrument performance ("instrument drift"), utilizing different LC columns and electrospray emitters. Often a biological study is executed by one group in one location at one time while the resulting samples are analyzed using LC/MS by an independent group in a separate location at a later time. The study designers are advised to include various quality control samples and use a complementary LC/MS sample queuing plan to guard the validity and objectivity of their study. Here, the important statistical principles are randomization, replication, and blocking, of which blocking is key.

A block is a set of samples spanning the interesting factors over which the nuisance factors are assumed to have a constant effect (the nuisance effect, however, may vary from block to block). Thus, block size, or number of samples in a block, is determined by the nuisance factor combinations. Consider a study investigating protein expression in diabetic tissue. Here, age, gender and body mass index (BMI) could be nuisance factors. A block in the diabetes design would be a sample from the diabetic tissue of interest with one combination of all nuisance factors-one tissue sample each from a non-diabetic, pre-diabetic and diabetic matched on age, gender and BMI. In its simplest form, a block is one replicate of the full biological design, or a complete mini-experiment containing one sample from each treatment combination. Blocking is quite common in biological studies and an experiment's blocks are the natural blocks for LC/MS processing. If the study design does not feature blocks, then study blocks solely for queuing LC/MS samples may be formed. Randomly select one sample from each treatment combination to fill a block. (Figure 4).

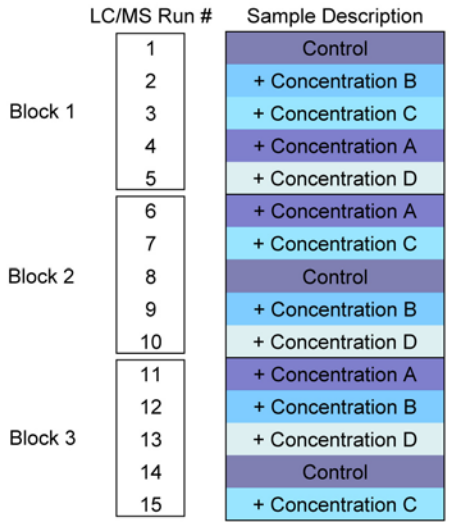

Figure 4. A simple example of a LC/MS queue.Shown is a partial LC/MS queue for an experiment where a cell line is exposed to 4 different concentrations of pesticides. Here a block contains 5 LC/MS injections one from each sample types-control, and those exposed to concentrations B, C, and D. Note that within each block the five different samples of peptides are in random order within each block.

The general stability of an LCMS processing line, LC column or MS instrument may be assessed with a controlled experiment featuring the sequential processing of numerous replicates of the same quality control sample across one or more processing lines. Here the objective is to identify the longest run of injections, or block size, over which the nuisance effects of an LCMS processing line are relatively constant. Suppose the LC/MS block size is larger than the study block size, then the study blocks effectively become the LC/MS blocks. That said, the samples within each existing study block should be randomly ordered, and then these blocks should be randomly queued for LC/MS processing. The aim of the LC/MS sample queuing plan is to control the confounding of nuisance LC/MS processing factors with the biological factors of interest (Figure 5). Though specific LC/MS nuisance factors are many, most can be sufficiently controlled by grouping under the major categorical variables: sample preparation set, LC column ID and LC/MS data acquisition start time.

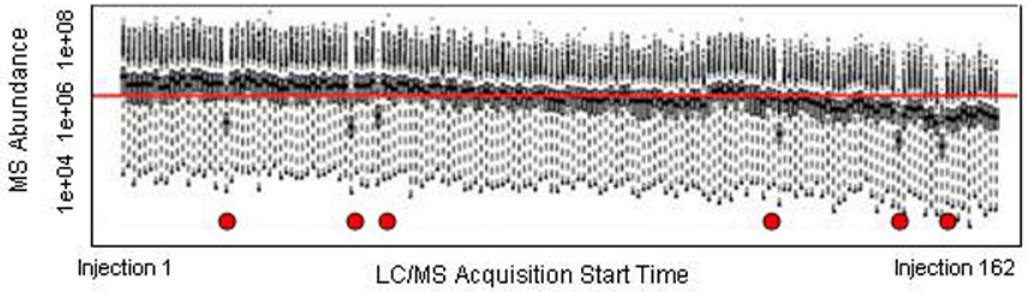

Figure 5. Advantages of proper LC/MS queuing.

A boxplot MS peptide abundances (y-axis) for 162 LC/MS injections of 27 separate total soluble digest of Trichoderma reesei. The red trend line is the median peptide abundance across all samples. Red dots are below six LC/MS injections that are well below the median observed for the remaining samples. All six are LC/MS technical replicates (injections) from a single bench top replicate, indicating a problem in the sample processing of this sample.

The mixed-effects model may also include terms that reflect a more complex design of the biological experiment, and terms that break out other LC/MS processing effects such as differences in sample preparations and time of MS acquisition. Terms not supported by measurements, such as peptides only observed in one LC/MS injection are excluded. The effectiveness of this modeling is limited by the amount and pattern of missing observations. In effect, only the information in the observed abundances is retained while the information in missing observations is discarded. This need not be the case. If the abundance data is converted to observation presence/absence, or 1/0, data, then the differences in the probability of a peptide observation across treatments may be modeled using additional statistical methods.

Overall the goal is to draw valid, objective, statistically defensible conclusions. As in the first look at the data, visual and tabular summaries are very effective. Here, however, the strength of the evidence need not be anecdotal because an appropriate analysis produces valid estimates of standard errors and confidence intervals. Careful interpretation, however, is required. It is important that the interpreter know the statistically valid interpretation of standard errors and confidence intervals. Each parameter estimate, or contrast of parameter estimates, has its own standard error and confidence interval.

6. Integrating proteomic data with other high throughput data

Technological advances in high-throughput technologies have been fueling a revolution in biology, enabling analyses of entire systems at a global scale (e.g., whole cells, tumors, or environmental communities). Thus far, the discussion has focused on the global profiling of proteins using high-throughput mass spectrometry (e.g., normalization approaches). In the context of systems biology this proteome information must be integrated with a plethora of additional information, both from other high-throughput "omic" technologies (e.g., transcriptomics and metabolomics), and supplementary information (e.g., functional annotations, cellular location predictions, regulatory elements). This task of data integration, with an eye on systems biology, requires multiple layers of computational tasks, including linking to data management systems, bioinformatics tools and statistical and visualization methods.

Data Management and Connectivity to Bioinformatics Tools

There are many challenges with managing data from heterogeneous data sources that range from simple access to the data to performing complex queries and workflows on the data to answer targeted questions of interest. In practice, the need to integrate and perform complex analyses on heterogeneous data sources results in ad hoc connections

between databases and software tools by writing small scripts, cutting and pasting queries, and basic manual labor. As a result, recent years have seen a surge in the development of new software tools focused on automating and simplifying these tasks. These bioinformatics resource tools tend to fall into four categories; semantic mapping, interoperation of heterogeneous bioinformatics databases, automated workflow analyses, and programs that integrate the data with bioinformatics software. Semantic mapping approaches, such as ToolBus (Eckart et al., 2003) and Taverna (Oinn; 2004), focus on defining translation engines that ensures that entities across environments are appropriately related. Alternatively, other approaches focus on the capability to access heterogeneous databases and merge directly on the data sources, such as BRIDGE (Goesmann et al., 2003). These works have led to subsequent tools, such as BioWarehouse (Lee et al., 2006) and GenFlow (Oikawa et al., 2004), that offer a combination of semantic mapping and database access capabilities. Alternatively, one can focus on the goal of the integration and define specialized workflows (Lu et al., 2006; Peleg and Altmann 2002). In some cases these methods are linked to statistical and visualization tools (Facius et al., 2005; Watson 2005).

Some current approaches, such as BRM (Shah et al., 2007), Gaggle (Shannon et al., 2006), and FACT (Kokocinski et al., 2005), focus on facilitating all of these capabilities (object mapping, database access, and generic workflows) into a single environment. Since both of these systems biology environments are built in Java, they can be easily installed and run by biologists and bioinformatics experts on publicly available websites: BRM (http://www.sysbio.org/dataresources/brm.stm) and Gaggle (http://gaggle.systemsbiology.net). These two integration and analysis tools have commonalities and the underlying programming languages allow them to work together. The BRM working environment (Figure 6) is given as an example of the multi-layer analyses allowed by these multi-capability software programs. On the top left is the project browser which allows the user to manage multiple heterogeneous datasets in a single space that gives information on each source, such as the number of rows and columns. The data set browser on the bottom allows the user to evaluate multiple data types in one view and codes each data source by color. To retrieve additional information associated with one or more data sources the data retrieval panel (top right) allows direct access to bioinformatics resources such as protein interactions (Bader et al., 2000; Xanarios 2000), pathways (Kanehisa and Goto 2000), and annotation data (Wheeler et al., 2006; Kersey et al., 2004; Bairoch et al., 2005). In addition, from this retrieval data panel, visualization softwares (Webb-Robertson et al., 2007; Shannon et al., 2003) associated with different types of data can be launch directly without any additional installations from the user.

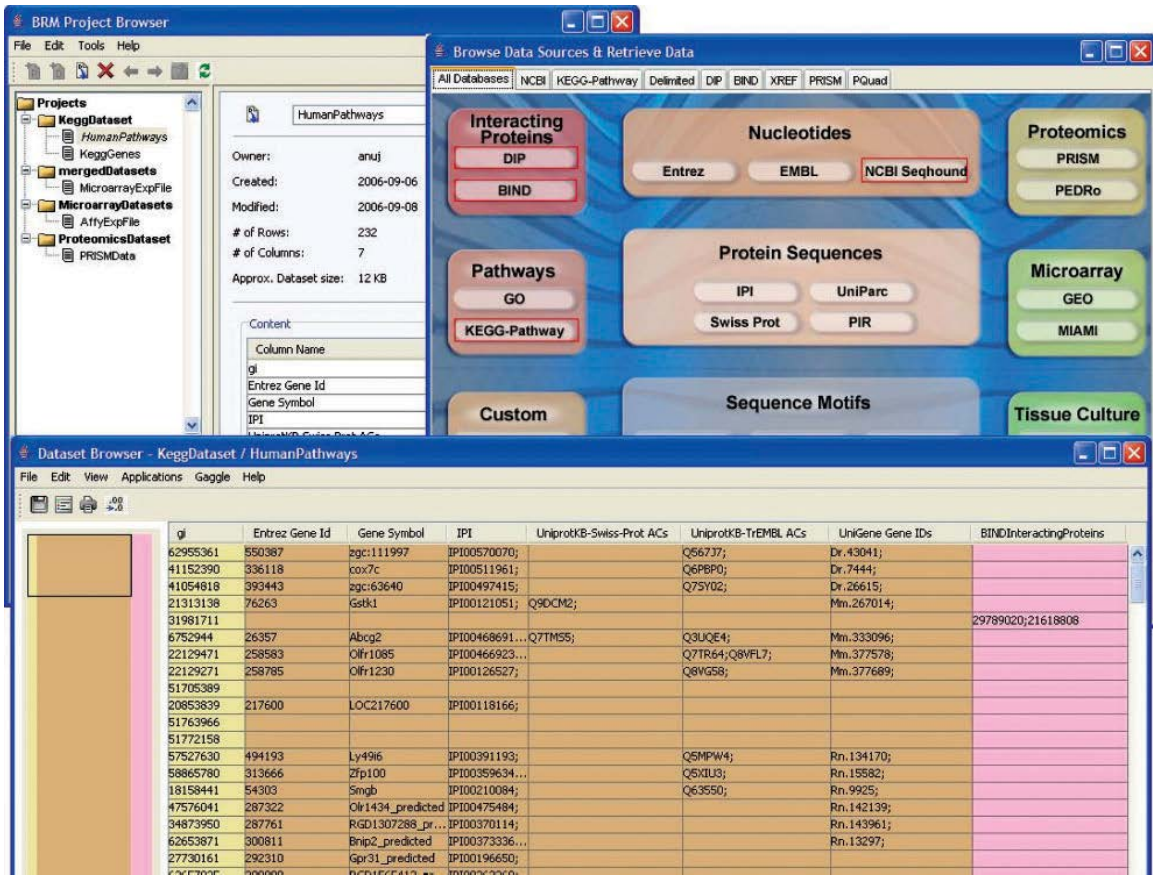

Figure 6: *A collage of the Bioinformatics Resource Manager (BRM) working environment, including the project browser (top left), data set browser (bottom), and data retrieval panel (top right) capabilities.*

Statistical Integration

There are many levels of integration that can be performed when evaluating multiple omics data sources as well as ancillary information (e.g., gene ontologies). The task is often complicated due to the heterogeneity of the data; for example, quantitative variables on multiple scale and categorical information. Many reviews have also been completed on integration of omics data focused towards systems biology (Aggarwal and Lee 2003; Reif et al., 2004; Nie et al., 2007). However, these reviews tend to focus on a generalized need and not the specifics of the statistical methods that may be employed. De Keersmaecker et al., (2006) nicely discusses the guiding principles of integration, such as balancing sensitivity and false discovery rates, global versus query specific analyses, supervised versus unsupervised methods, and sequential versus concurrent methods.

In general, statistical methods tend to fall into two categories, unsupervised (exploratory) analyses or supervised learning. Unsupervised methods, such as principal component analysis (PCA) (Johnson and Wichern 1992), optimize some feature of the data, such as variance, which may reveal clustering tendencies of the data in a lower dimensionality. These methods are for exploration purposes to try to identify underlying structure in the data. Alternatively, supervised learning assumes that the response is known, or has been

measured, and the goal is to find a correlative model between the set of features and the response, such as with regression (Neter et al., 1996). These methods are predictive in the sense that if one attains the set of features for a new observation the response can be predicted from the model.

In respect to statistical data integration, irrelevant to the actual statistical model employed, there are generally three basic approaches to merge the data for statistical analysis, which is highly dependent upon the type of data that is being considered. The first is feature integration where the individual datasets are merged into a global dataset and then evaluated using either supervised or unsupervised learning. The second is to individually evaluate each dataset with methods, such as clustering, and then statistically merge the results. The last method is to transform each dataset into an alternate representation, such as a network or kernel, and merge the data in this new dimensional space: these methods are usually used in conjunction with supervised learning. These three strategies are described briefly, as well as the benefits and caveats of each approach.

Feature Integration

One of the most common approaches in data integration is simply feature integration. If dataset *A* consists of *m* features, $D_A = [f_1, f_2, \cdots f_m]$, and dataset *B* consists of *n* features, $D_B = [g_1, g_2, \cdots g_n]$, than the integrated dataset is simply:

$$D_{AB} = [f_1, f_2, \cdots f_m, g_1, g_2, \cdots g_n]. \quad (1)$$

This can be performed by either be by merging the two datasets such that each observation in one dataset matches that in another, e.g., each protein corresponds to a gene, or by attempting to separate specific events, such as the toxicity of a compound so that the biological samples are the observations and the biomolecular molecules are the features. The task of merging data for the goal of integrating microarray and proteomic data has been reviewed (Waters et al., 2006) and can be accomplished using tools such as those previously described. Additionally, Cox et al. (2005) give a review of clustering and correlation based approaches for merged datasets. This approach is only feasible when the variables are of a common type, e.g., qualitative, since normalization is typically a necessity to put each variable on the same scale. However, given the appropriate scaled dataset, most multivariate statistical methods, such as clustering or regression, could be employed to analyze trends or relationships in the data. In the field of proteomics, this approach is most commonly used to merge ancillary information with peptide identification results to improve the quality of the results, i.e., improve sensitivity (Anderson et al., 2003; Cannon et al., 2005). These approaches describe a peptide as a set of disparate features associated with identification metrics, such as the cross correlation score from SEQUEST (Eng et al., 1994) and fraction of matched peaks, and use the supervised learning algorithm support vector machine (SVM) (Cristianini and Shawe-Taylor 2000; Vapnik 1995) to determine correct from incorrect identifications.

The primary caveats of feature integration are the need for a one-to-one correspondence between objects in each dataset and that variable types, such as categorical information, are difficult to merge. In addition, the contribution of each feature is often not evident but of interest. Thus, feature selection methods typically follow the initial analyses.

Overall, with feature integration, care must be taken to assure that the data are appropriate to merge and properly normalized.

Individual Analyses Followed by Integration

An alternative approach to integration is to evaluate each dataset individually and then merge the results of each analysis. An unsupervised approach to this task is to cluster each dataset into some number of clusters and then merge the clusterings, commonly referred to as metaclustering (Topchy et al., 2004; Zeng et al., 2002). These methods have been demonstrated to be of use in biology (Barutcuoglu et al., Kano et al., 2005; Kasturi and Acharya 2004) If each observation is treated as a probability of being associated with a specific cluster than a simple Bayesian approach can be taken to merge these results. Two primary benefits of this approach is the capability to integrate multiple data formats, e.g., qualitative and quantitative, and the low dimensionality of the results are conducive to visualization (Havre et al., 2005). Figure 7 gives an example where three types of experimental data (Powerblot, FTICR proteomics, and microarray) over a time course. For each dataset at each time point the dataset are clustered into three classes (up-regulated, down-regulated, neither). The top and bottom axes demonstrate which colored lines belong to each data type and their respective results at each time point. As seen in the figure, common trends among the three datasets can be easily observed. In addition, the bottom metaclustering is a merged result over the entire time course to highlight statistical trends among the data.

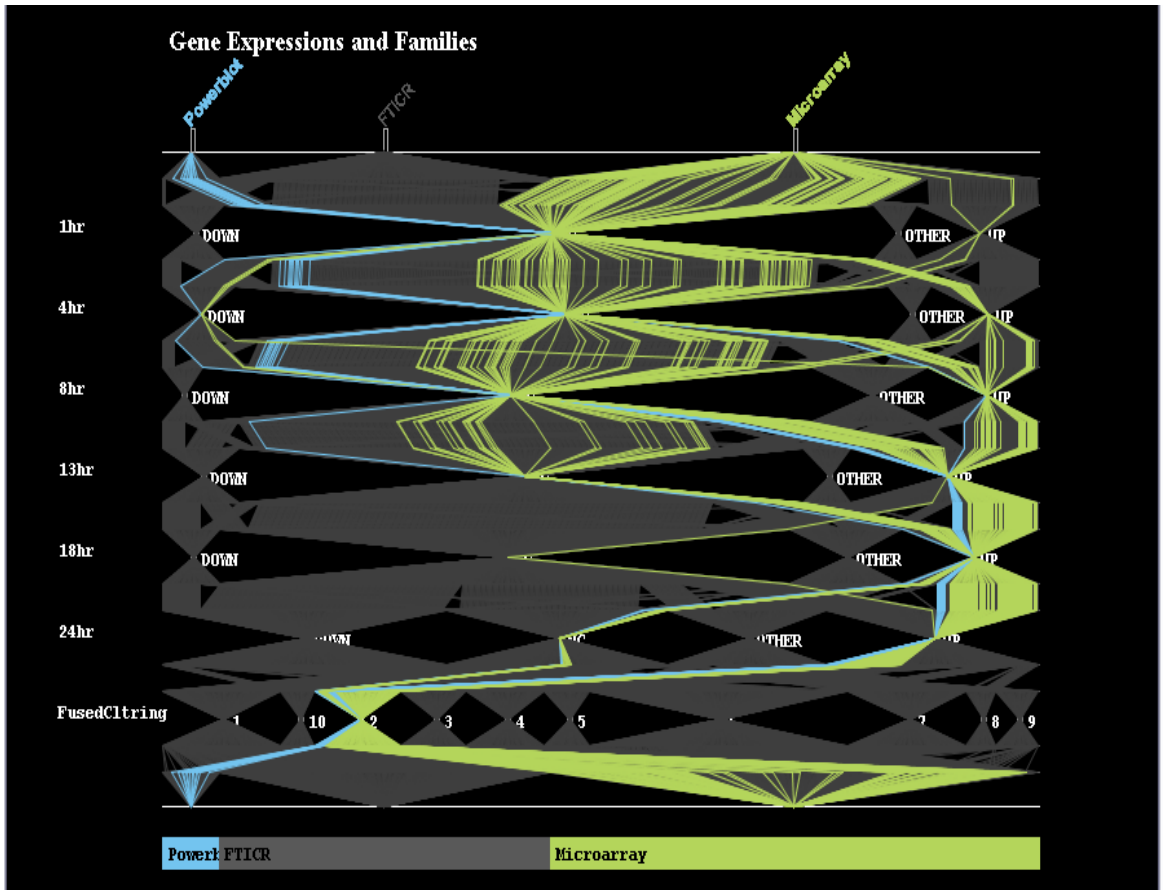

Figure 7: *The Juxter visualization tool demonstrates the capability to integrate invidual datasets, in this case Powerblot, FTICR, and Microarray data, each at an individual time point. The top and bottom tiers represent the data type which each layer in the visualization shows which genes or proteins fall into the categories of up-, down-, or non-regulated at each time point. The last layer in the visualization gives the statistically merged results over the entire time course.*

Alternative to classifying or clustering the data, as seen in Figure 2, often in proteomics and biology, the end goal is to find a set of biomolecules that are relevant to the question of interest, e.g., which proteins are associated with a virulent versus non-virulent pathogen. In this case researchers commonly use statistical tests of significance and assign p-values, or normalized false discovery rates, to each biomolecule. Thus, all the datasets can be reduced down to a set of p-values, which can then be merged into a level of significance associated with related entities. Recent years have seen these methods become much more robust by accounting for biological nuances and using multiple statistics to evaluate the significance of individual biolomolecular species (e.g., genes, proteins, metabolites). Fagan et al., (2007) first evaluate each dataset using PCA to visualize relationships in each data source and reduce the dimensionality of the integration task. They then use co-inertia analysis to evaluate correlations across thedatasets, which better accounts for biological issues that arise in direct correlation analyses because of post-transcriptional and -translational regulations. POINTILLIST (Hwang et al., 2006) uses a weighted version of several statistical metrics of significance to derive a network model where the integrated p-value measure indicates the degree of confidence in a node or edge being a true component of the system of interest where a node represents a biomolecular species.

The benefit of these methods, both integration of clusters and statistical levels of significance, is that they can better handle datasets of vastly different sizes and types. Additionally, normalization only needs to be performed within each dataset. Even further, as seen in Figure 2, there doesn't need to be a one-to-one mapping between datasets under the condition that clusterings are the end goal. Some of the statistical significance integration approaches can account for missing data, a common problem in proteomics. The caveats is that in many cases since there may not be a one-to-one mapping between the datasets, interpretation may be difficult and time-consuming.

Integration in Feature Space via Data Transformation

In supervised learning it is often the case that the data is transformed into an alterative representation, such as a relationship or a kernel matrix. In biology, often data is represented as the relationship between biomolecular entities, for example correlations between genes that might relate to a common regulation or links between proteins that represent possible interactions. These relationship matrices can be merged into a more accurate view of the system using methods such as Bayesian networks where the relationship matrices are the input (Gilchrist et al., 2004; Huttenhower and Troyanskaya 2006; Troyanskaya et al., 2003). This approach is slightly different from that described above since the relationship matrix itself is not typically the final result for an individual dataset, but an intermediate representation used for the task of learning. Since the data

are merged at an intermediate form a major benefit of this approach is that the data do not have to have a one-to-one mapping. The largest caveat is that these methods are often computationally intensive in learning the parameters of the model.

A more abstract approach to integration of transformed data is kernel fusion. A kernel function is a transformed projection of the data that in principle enhances linear separability. Kernel functions are especially powerful for dataset that are not linearly separable by mapping the data into a space that can be linearly separated by a SVM. Individual kernel functions for each dataset can be merged into an integrated kernel (Lanckriet et al., 2004),

$$K_{Int} = \mu_1 K_1 + \mu_2 K_2 + \cdots + \mu_n K_n, \quad (2)$$

where K_i is the kernel associated with the *i*-th dataset. K_{Int} can be used to build a supervised model in the same manner as for a single data source. This is a very powerful statistical approach, however it has the same caveats as the feature integration method; a onoe-to-one mapping between biomolecular entities. However, with this approach it is much easier to integrate information from other computational tools, such similarity between entities by protein domains or sequence similarity.

7. Summary

In 1958 Francis Crick laid out the "Central Dogma" of biology:

> "…once 'information' has passed into protein *it cannot get out again.* In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the *precise* determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein." (Crick 1958)

In other words, proteins are an endpoint for the information encoded within the genome. Thus, they should and do merit a strong research focus. The structure of proteins, however, makes there study at a global scale much more difficult than that of nucleic acid for which high throughput sequencing and hybridization approaches exist. The study of all of the proteins encoded by the genome, or the "proteome" (the term coined by Australian researcher Marc Wilkins in 1994) relies on protein or peptide separation followed by mass spectrometry. This approach has proven to be the most efficient method to identify protein sequences *en masse*. However, given the complex mixtures and the nature of the approach, there are several caveats associated with modern proteomics.

We have explored in this chapter many caveats associated with current proteomics methods. Gene models are crucial to proteomics, because they are used as the basis for database searching algorithms that are used to match mass spectra generated by global proteomics. Incorrect models leads to both false positive and false negative peptide sequence information. Processing of samples for proteomic analysis is also important –

developing a method for protein isolation can be generalized, but with the breadth of organisms being studied, sample cleanup can vary. Also crucial to processing, is the overall experimental design. A statistically rigorous design is integral to downstream analysis and can be useful in identifying samples that fail due to processing or instrument error. The correct design of an experiment also translates to the ability to apply statistical modeling approaches that move proteomics from a qualitative method to a quantitative method. Finally, tools and methods for the integration of proteomic and other high throughput global analyses such as microarrays and metabolomics, are needed because the proteome is only one tool of several that are needed to build a hypotheses and models of biological systems.

The proteomics field continues to advance rapidly. Looking forward, we will need further advances in gene modeling, mass spectrometry, rigorous statistical approaches and bioinformatics tools for proteomics to have the robustness of methods currently in use for genome sequencing and transcriptome analysis. Event though there will be improvements and refinements moving forwards, there is still much that we can learn using currently available approaches to proteome analysis.

References

K. Aggarwal and K.H. Lee, Functional genomics and proteomics as a foundation for systems biology. *Brief. Funct. Genomic. and Proteomic.*, **2**, 175-184 (2003).

J.E. Allen, M. Pertea and S.L. Salzberg, Computational gene prediction using multiple sources of evidence, *Genome Res*., **14**, 142-148 (2004).

J.E. Allen and S.L. Salzberg, JIGSAW: integration of multiple sources of evidence for gene prediction, *Bioinformatics*, **21**, 3596-3603 (2005).

S.F. Altschul and E.V. Koonin, Iterated profile searches with PSI-BLAST-a tool for discovery in protein databases, *Trends Biochem. Sci*., **23**, 444-447 (1998).

D.C. Anderson, W. Li , D.G. Payan and W.S. Noble, A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J. Proteome Res.*, **2**, 137-146 (2003).

M. Ashburner, C.A Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin and G. Sherlock, Gene ontology: tool for the unification of biology, *Nat. Genet*., **25**, 25-29 (2000).

G.D. Bader, D. Betel, and C.W. Hogue, BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **31**, 248-250 (2003).

A. Bairoch, R. Apweiler, C.H. Wu, W.C.Barker, B.Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi and L.S. Yeh, The Universal Protein Resource (UniProt). *Nucleic Acids Res*, **33**, D154-D159 (2005).

M. Bantscheff, M. Schirle, G. Sweetman, J. Rick and B. Kuster, Quantiative mass spectrometry in proteomics: a critical review, *Anal. Bioanal. Chem*, **389**, 1017-1031 (2007).

T. Barrett, T.O. Suzek, D.B. Troup, S.E. Wilhite, W.C. Ngau, P. Ledoux, D. Rudnev, A.E. Lash, W. Fujibuchi and R. Edgar, NCBI GEO: mining millions of expression profiles- database and tools, *Nucleic Acids Res*., **33,** D562-566 (2005).

Z. Barutcuoglu, R.E. Schapire and O.G. Troyanskaya, Hierarchical mulit-label prediction of gene function. *Bioinformatics*, **22**, 830-836 (2006).

G. Bestel-Corre, E. Dumas-Gaudot and S. Gianinazzi, Proteomics as a tool to monitor plant-microbe endosymbioses in the rhizosphere, *Mycorrhiza*, **14**, 1-10 (2004).

E. Birney, M. Clamp and R. Durbin, GeneWise and genomewise, *Genome Res*., **14**, 988-995 (2004).

A. Bodzon-Kulakowska, A. Bierczynska-Krzysik, T. Dylag, A. Drabik, P. Suder, M. Noga, J. Jarzebinska. and J. Silberring, Methods for sample preparation in proteomic research, *J. Chromatogr. B*, **849**, 1-31 (2007).

B.R. Braun, M. van Het Hoog, C. d'Enfert, M. Martchenko, J. Dungan, A. Kuo, D.O. Inglis, M.A. Uhl, H. Hogues, M. Berriman, M. Lorenz, A.Levitin, U. Oberholzer, C. Bachewich, D. Harcus, A. Marcil, D. Dignard, T. Iouk, R. Zito, F. Tekaia, K. Rutherford, E. Wang, C.A. Munro, S. Bates, N.A. Gow., L.L. Hoyer, G. Kohler, J. Morschhauser, G. Newport, S. Znaidi, M. Raymond, B. Turcotte, G. Sherlock, M. Costanzo, J. Ihmels, J. Berman, D. Sanglard, N. Agabian, A.P. Mitchell, A.D. Johnson, M. Whiteway and A. Nantel, A human-curated annotation of the *Candida albicans* genome, *PLoS Genet*. **1**, 36-57 (2005).

M. Burset and R. Guigo, Evaluation of gene structure prediction programs, *Genomics*, **34**, 353-67 (1996).

B. Canas, C. Pineiro, E. Calvo, D. Lopez-Ferrer and J.M. Gallardo, Trends in sample preparation for classical and second generation proteomics, *J. Chromatogr. A*., **1153**, 235-258 (2007).

W. R. Cannon, K.H. Jarman, B.J. Webb-Robertson, D.J. Baxter, C.S. Oehmen, K.D. Jarman, A. Heredia-Langner, K.J. Auberry and G.A. Anderson, Comparison of probability and likelihood models for peptide identification from tandem mass spectrometry data. *J. Proteome Res.*, **4**, 1687-1698 (2005).

T.J. Carver, K.M. Rutherford, M. Berriman, M. Rajandream, B.G. Barrell and J. Parkhill. ACT: the artemis comparison tool. *Bioinformatics* **21**, 3422-3423 (2005).

R.B. Cole, Some tenets pertaining to electrospray ionization mass spectrometry, *J. Mass Spectrom.*, **35**, 763-772 (2000).

T.P. Conrads, G.A. Anderson, T.D. Veenstra, L. Pasa-Tolic and R.D. Smith, Utility of Acurate Mass Tags for Proteome-Wide Protein Identification, Anal. Chem., 72, 3349-3354 (2000).

F,H. Crick, On protein synthesis, Symp. Soc. Exp. Bio., 12, 138-163 (1958).

B. Cox, T. Kislinger and A. Emili, Integrating gene and protein expression data: pattern analysis and profile mining. *Methods*, **35**, 303-314 (2004).

R. Craig and R.C. Beavis, A method for reducing the time required to match protein sequences with tandem mass spectra, *Rapid Commun. Mass Spectrom.*, **17**, 2310-2316 (2003).

N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge, 2000.
D.S. Daly, K.K. Anderson, E.A. Panisko, S.P. Purvine, R. Fang, M.E. Monroe and S.E. Baker, Mixed-effects statistical model for comparative LC-MS proteomics studies, *J. Proteome Res.*, In press, epub pr070441I (2008).

S.E. De Keersmaecker, I.M. Thijis, J. Vanderlevden and K. Marchal, Integration of omics data: how well does it work for bacteria? *Mol. Microbiol.*, **62**, 1239-1250 (2006).

C. Dewey, J.Q. Wu, S. Cawley, M. Alexandersson, R. Gibbs and L. Patcher, Accurate identification of novel human genes through simultaneous gene prediction in human, mouse, and rat, *Genome Res.*, **14**, 661-664 (2004).

B. Dujon, D. Sherman, G. Fischer, P. Durrens, S. Casaregola, I. Lafontaine, J. DeMontigny, C. Marck, C. Neuveflise, E. Talla, N. Goffard, L. Frangeul, M. Aigle, V. Anthouard, A. Babour, V. Barbe, S. Barnay, S. Blanchin, J.M. Beckerich, E. Beyne, C. Bleykasten, A. Boisrame, J. Boyer, L. Cattolico, F. Confanioleri, A. De Daruvar, L. Despons, E. Fabre, C. Fairhead, H. Fery-Dumazet, A. Groppi, F. Hantraye, C.Hennequin, N. Jauniaux, P. Joyet, R. Kachouri, A. Kerrest, R. Koszul, M. Lemaire, I. Lesur, L. Ma, H. Muller, J.M. Nicaud, M. Nikolski, S. Oztas, O. Ozier-Kalogeropoulos, S. Pellenz, S. Potier, G.F. Richard, M.L. Straub, A. Suleau, D. Swennen, F. Tekaia, M. Wesolowski-Louvel, E. Westhof, B. Wirth, M. Zeniou-Meyer, M. Bolotin-Fukuhara, A. Thierry, C. Bouchier, B. Caudron, C. Scarpelli, C. Gaillardin, J. Weissenbach, P. Wincker and J.L. Souciet, Genome evolution in yeasts, *Nature*, **430**, 35-44 (2004).

J.D. Eckart and B.W. Sobral, A life scientist's gateway to distributed data management and computing: the PathPort/ToolBus framework. *Omics*, **7**, 79-88 (2003).

J.K. Eng, A.L. McCormack. and J.R.Yates III, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, **5**, 976-989 (1994).

O. Emanuelsson, S. Brunak, G. von Heijne and H. Nielsen, Locating proteins in the cell using TargetP, SignalP, and related tools, *Nat. Protoc*., **2**, 953-971 (2007).

A. Facius, C. Englbrecht, F. Birzele, A. Groscurth, B. Schmid, S. Wanka and W. Mewes, PRIME: A graphical interface for interrogating genomic/proteomic databases. *Proteomics*, **5**, 76-80 (2005).

A. Fagan, A.C. Culhane and D.G. Higgins, A multivariate analysis approach to the integration of proteomic and gene expression data. *Proteomics*, 7: 2162-2171 (2007).

S.B. Ficarro, M.L. McCleland, P.T. Stukenberg, D.J. Burke, M.M. Ross, J. Shabanowitz, D.F. Hunt and F.M. White, Phophoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*, *Nat Biotech.*, **20**, 301-305 (2002).

R.D. Fleischmann, M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick, et al., Whole-genome random sequencing and assembly of *Haemophilus influenzaie* Rd, *Science*, **269**, 496-512 (1995).

K. Gevaert and J. Vandekerckhove, Protein identification methods in proteomics, *Electrophoresis*, **21**, 1145-1154 (2000).

M.A. Gilchrist, L.A. Salter and A. Wagner, A statistical framework for combining and interpreting proteomic datasets. *Bioinformatics*, **20**, 689-700 (2004).

A. Goesmann, B. Binke, O. Rupp, L. Krause, D. Bartels, M. Dondrup, A.C. McHardy, A. Wilke, A. Pühler and F. Meyer, Building a BRIDGE for the integration of heterogeneous data from functional genomics into a platform for systems biology. *J. Biotechnol.*, 106, 157-167 (2003).

A. Goesmann, B. Linke, D. Bartels, M. Dondrup, L. Krause, H. Neuweger, S. Oehm, T. Paczian, A. Wilke and F. Meyer, BRIDEP – the BRIDGE-based genome-transcriptome-proteome browser. *Nucleic Acids Res.*, 33, W710:W716 (2005).

A. Goffeau, B.G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J.D. Hoheisel, C. Jacq, M. Johnston, E.J. Louis, H.W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin and S.G. Oliver, Life with 6000 genes, *Science*, **274**, 563-567 (1996).

M.B. Goshe, T.P. Conrads, E.A. Panisko, N.H. Angell, T.D. Veenstra and R.D. Smith, Phosphoprotein isotope-coded affinity tag approach for isolating and quantitating phosphopeptides in proteome-wide analyses, *Anal. Chem*, **73**, 2578-2586 (2001).

W.J. Griffiths, A.P. Jonsson, S. Liu, D.K. Rai and Y. Wang. Electrospray and tandem mass spectrometry in biochemistry. *Biochem. J.* **355**, 545-561 (2001).

R. Guigo, P. Flicek, J.F. Abril, A. Reymond, J. Lagarde, F. Denoeud, S. Antonarakis, M. Ashburner, V.B. Bajic, E. Birney, R. Castelo, E. Eyras, C. Ucla, T.R. Gingeras, J. Haroow, T. Hubbard, S.E. Lewis and M.G. Reese, EGASP: the human ENCODE genome annotation assessment project, *Genome Biol*., **7**, Suppl 1, S2.1-31 (2006).

M.G. Reese, G. Hartzell, N.L Harris, U. Ohler, J.F. Abril. and S.E. Lewis, Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.*, 10, 483-501 (2000).

S.P. Gygi, B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb and R. Aebersold, Quantiative analysis of complex protein mixtures using isotope-coded affinity tags, *Nat. Biotechnol.*, **17**, 994-999 (1999).

B.J. Haas, A.L. Delcher, S.M. Mount, J.R. Wortman, R.K. Smith Jr., L.I. Hannick, R. Maiti, C.M. Ronning, D.B. Rusch, C.D. Town, S.L. Salzberg and O. White, Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies, *Nucleic Acids Res.*, **31**, 5654-5666 (2003).

S.L. Havre, B.M. Webb-Robertson, A. Shah, C. Posse, B. Gopalan and F.J. Brockman, Bioinformatic insights from metagenomics through visualization. *Proceedings of the IEEE Computational Systems Bioinformatics Conference*, 341-350 (2005).

C. Huttenhower and O.G. Troyanskaya, Bayesian data integration: a function perspective. *Computational Systems Bioinformatics Conference*, 341-351 (2006).

D. Hwang, A.G. Rust, S. Ramsey, J.J. Smith, D.M. Leslie, A.D. Weston, P. de Atauri, J.D. Aitchison, L. Hood, A.F. Siegel and H. Bolouri, A data integration methodology for systems biology *PNAS*, **102**, 17296-17301 (2005).

O. Jaillon, J.M. Aury, B. Noel, A. Policriti, C. Clept, A. Casagrande, N. Choisne, N. Vitulo, C. Jubin, A. Vezzi, F. Legeai, P. Hugueney, C. Dasilva, D. Horner, E. Mica, D. Jublot, J. Poulain, C. Bruyere, A. Billault, B. Segurens, M. Gouyvenoux, E. Ugarte, F. Cattonaro, V. Anthouard, V. Vico, C. Del Fabbro, M. Alaux, G. Di Gaspero, V. Dumas, N. Felice, S. Paillard, I. Juman, M. Moroldo, S. Scalabrin, A. Canaguie, I. Le Clainche, G. Malacrinda, E. Durand, G. Pesole, V. Laucou, P. Chatlet, D. Merdinoglu, M. Delledonne, M. Pezzotti, A. Lechamy, C. Scarpelli,F. Artiguenave, M.E. Pe, G. Valle, M. Morgante, M. Caboche, A.F. Adam-Blondon, J. Weissenbach, F. Quetier and P. Wincker; French-Italian Consortium for Grapevine Genome Characterization, The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla, *Nature*, **449**, 463-467 (2007).

R. A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis*. 3rd ed., Prentice Hall, Upper Saddle River, 1992.

M. Kanehisa and S.Goto, KEGG: Kyoto Encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27-30 (2000).

M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno and M. Hattori, The KEGG resource for deciphering the genome, *Nucleic Acids Res.*, **32**, D277-80 (2004).

M. Kano, S. Tsutsumi, N. Kawahara, Y. Wang, A. Mukasa, T. Kirino and H. Aburatani, A meta-clustering analysis indicates distinct pattern alteration between two series of gene expression profiles for induced ischemic tolerance in rats. *Physiol. Genomics*, **21**, 274-283 (2005).

J,. Kasturi and R. Acharya, Clustering of diverse genomic data using information fusion. *Bioinformatics*, **21**, 423-429 (2004).

P.J. Kersey, J. Duarte, A. Williams, Y. Karavidopoulou, E. Birney and R. Apweiler, R. The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, **4**, 1985-1988 (2004).

G.R. Kiebel, K.J. Auberry, N. Jaitly, D.A. Clark, M.E. Monroe, E.S. Peterson, N. Tolic, G.A. Anderson and R.D. Smith, PRISM: a data management system for high-throughput proteomics, *Proteomics*, **6**, 1783-1790 (2006).

F. Kokocinski, N. Delhomme, G. Wrobel, L. Hummerich, G. Toedt and P. Lichter P. FACT – a framework for the functional interpretation of high-throughput experiments. *BMC Bioinformatics*, **6**, 161 (2005).

E.V. Koonin, N.D. Fedorova, S.D. Jackson, A.R. Jacobs, D.M. Krylov, K.S. Makarova, R. Mazumder, S.L. Mekhedov, A.N. Nikolskaya, B.S. Rao, I.B. Rogozin, S. Smirnov, A.V. Sorokin, A.V. Sverdlov, S. Vasudevan, Y.I. Wolf, J.J. Yin and D.A. Natale, A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes, *Genome Biol.*, **5**, R7 (2004).

I. Krof, Gene finding in novel genomes, *BMC Bioinformatics*, **14,** 59 (2004).

N.M. Laird and J.H. Ware, Random effects models for longitudinal data, *Biometrics*, **38**, 963-974 (1982).

G.R. Lanckriet, T. De Bie, N. Cristianini, M.I. Jordaan and W.S. Noble, A statistical framework for genomic data fusion. *Bioinformatics*, **20**, 2625-2635 (2004).

T. J. Lee, Y. Pouliot, V. Wagner, P. Gupta, D.W. Stringer-Calvert, J.D. Tenenbaum and P.D. Karp, BioWarehouse: a bioinformatics database warehouse toolkit. *BMC Bioinformatics*, 7, 170 (2006).

Q. Lu, P. Hao, V. Curcin, W. He, Y.Y. Li, Q.M. Luo, Y.K. Guo Y.K. and Li Y.X., KDE Bioscience: platform for bioinformatics analysis workflows. *J. of Biomed. Inform.*, 39, 440-450 (2005).

A.V. Lukashin and M. Borodovsky, GeneMark.hmm: new solutions for gene finding, *Nucleic Acids Res*., **26,** 1107-1115 (1998).

W.H. Majoros, M. Pertea and S.L. Salzberg, Efficient implementation of a generalized pair hidden Markov model for comparative gene finding, *Bioinformatics,* **21,** 1782-1788 (2005).

M. Mann and M. Wilm, Error tolerant identification of peptides in sequence databases by peptide sequence tags, *Anal. Chem.,* **66**, 4390-4399 (1994).

F. Martin, A. Aerts, D. Ahrén, A. Brun, E.G.J. Danchin, F. Duchaussoy, J. Gibon, A. Kohler, E. Lindquist, V. Pereda, A. Salamov, H.J. Shaprion, J. Wuyts, D. Blaudez, M. Buée, P. Brokstein, B. Canbäck, D. Cohen, P.E. Courty, P.M. Coutinho, C. Delaruelle, J.C. Detter, A. Deveau, S. DiFazio, S. Duplessis, L. Fraissinet-Tachet, E. Lucic, P. Frey-Klett, C. Fourrey, I. Feussner., G. Gay, J. Grimwood, P.J. Hoegger, P. Jain, S. Kilaru, J. Labbé, Y.C. Lin, V. Legué, F. Le Tacon, R. Marmeisse, D. Melayah, B. Montanini, M. Muratet, U. Nehls, H. Niculita-Hirzel, M.P. Oudot-Le Secq, M. Peter, H. Quesneville, B. Rajashekar, M. Reich, N. Rouhier, J. Schmutz, Y. Yin, M. Chalot, B. Henrissat, U. Kües, S. Lucas, Y. Van de Peer, G.K. Podila, A. Polle, P.J. Pikkila, P.M. Richardson, P. Rouzé, I.R. Sanders, J.E. Stajich, A. Tunlid, G. Tuskan and I. V. Grigoriev The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis *Nature*, in press (2008).

R. Marouga, S. David and E. Hawkins. The development of the DIGE system: 2D fluorescence difference gel analysis technology, *Anal. Bioanal. Chem*., **382**, 669-678 (2005).

M.L. Medina, P.A. Haynes, L. Breci and W.A. Francisco, Analysis of secreted proteins from *Aspergillus flavus*, *Proteomics*, **5**, 3153-3161 (2005).

M.L. Medina, U.A. Kiernan and W.A. Francisco, Proteomic analysis of rutin-inducted secreted proteins from *Aspergillus flavus*, *Fungal Genet. Biol*., **41**, 327-335 (2004).

A.I. Nesvizhskii, O. Vitek and R. Abersold, Analysis and validation of proteomic data generated by tandem mass spectrometry, *Nat. Methods*, **4**, 787-798 (2007).

J. Neter, H. Kutner, C.J. Nachtsheim and W. Wasserman, *Applied Linear Regression Models*, 3rd ed., McGraw-Hill Companies, Inc., Chicago, 1996.

L.Nie, G. Wu, D.E. Culley, J.C. Scholten and W. Zhang, Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications. *Crit. Rev. Biotechnol.*, **27**, 63-75 (2007).

NIST/SEMATECH., *e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook, 2008.

M.K. Oikawa, M.E.B. Broinizi, A. Dermargos, A.H. Armelin and J.E. Ferreira, GenFlow: Generic flow for integration, management and analysis of molecular biology data. *Genet. Mol. Biol.*, **27**, 691-695 (2004).

T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Geenwood, T. Carver, K. Glover, M.R. Pocock, A. Wipat and P. Li, Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20, 3045-3054 (2004).

S.E. Ong, B. Blagoev, I. Kratchmarova, D.B. Kristensen, H. Steen, A. Pandey and M. Mann, Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics, *Mol. Cell Proteomics*, **1**, 376-386 (2002).

S.E. Ong, J.F. Foster and M. Mann, Mass spectrometric-based approaches in quantitative proteomics, *Methods*, **29**, 124-130 (2003).

H. Parkinson, U. Sarkans, M. Shojatalab, N. Abeyguawardena, S. Contrino, R. Coulson, A. Farne, G.G. Lara, E. Holloway, M. Kapushesky, P. Lilja, G. Mukherjee, A. Oezcimen, T. Rayner, P. Rocco-Serra, A. Sharma, S. Sanson and A. Brazma, ArrayExpress- a public repository for microarray gene expression data at the EBI, *Nucleic Acids Res.*, **33**, D554-555 (2005).

H.D. Patterson and R. Thompson, Recovery of inter-block information when block sizes are unequal, *Biometrika*, **58**, 545-554 (1971).

V. Pavlovic, A. Garg and S. Kasif, A Bayesian framework for combining gene predictions, *Bioinformatics*, **18**, 19-27 (2002).

M. Peleg, I. Yeh and R.B. Altman,Modelling biological processes using workflow and Petri Net models. *Bioinformatics*, **18**, 825-837 (2002).

J.C. Pinheiro and D.M. Bates, *Mixed-effects models in S and S-plus*, Springer-Verlag, New York, 2000.

S.C. Potter, L. Clarke,V. Curwen, S. Keenan, E. Mongin, S.M. Searle, A. Stabenau, R. Storey and M. Clamp, The Ensembl analysis pipeline, *Genome Res., 14*, 935-941 (2004).

S.O. Purvine, A.F. Picone and E. Kolker, Standard mixtures for proteomics, *OMICS*, **8**, 79-92 (2004).

P.G. Righetti, A. Castagna, F. Antonucci, C. Piubelli, D. Cecconi, N.P. Campostrini, P. Antonioli, H Astner and M. Hamdam, Critical survey of quantitative proteomics in two-dimensional electrophoretic approaches, *J. Chromatogr. A.*, 1051, 3-17 (2004).

M.G. Reese, G. Hartzell, N.L. Harris, U. Ohler, J.F. Abril and S.E. Lewis, Genome annotation assessment in *Drosophila melanogaster*, *Genome Res*. **10**, 482-501 (2000).

D.M. Reif, B.C. White and J.H. Moore, Integrated analysis of genetic, genomic and proteomic data. *Expert Rev. Proteomics*, **1**, 67-75 (2004).

P.L. Ross, Y.N. Huang, J.N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson and D.J. Pappin, Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents, *Mol. Cell. Proteomics*, **3**, 1154-1169 (2004).

A.A. Salamov and V.V. Solovyev, *Ab initio* gene finding in *Drosophila* genomic DNA, *Genome Res.*, **10**, 516-522 (2000).

A. Schiex, A. Moisan and P. Rouze, EuGene: an eukaryotic gene finder that combines several sources of evidence, *Computational Biology*, Springer, Heidlberg, 2001.

S.R. Searle, G. Casella and C.E. McCulloch, *Variance components*, John Wiley & Sons, New York, 1992.

A.R. Shah, M. Singhal, K.R. Klicker, E.G. Stephan, H.S. Wiley and K.M. Waters, Enabling high-throughput data management for systems biology: the Bioinformatics Resource Manager. *Bioinformatics*, **23**, 906-909 (2007).

P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13, 2498-2504 (2003).

P.T. Shannon, D.J. Reiss, R. Bonneau and N.S. Baliga, The Gaggle: an open-source software system for integrating bioinformatics software and data sources. *BMC Bioinformatics*, 7, 176 (2006).

P.K. Smith, R.I. Krohn, G.T. Hermanson, A.K. Mallia, F.H. Gartner, M.D. Provenzano, E.K. Fujimoto, N.M. Goeke, B.J. Olson and D.C. Klenk, Measurement of protein using bicinchoninic acid, *Anal. Biochem*., **150**, 76-85 (1985).

Solovyev V.V., Structure, properties and computer identification of eukaryotic genes, *Bioinformatics from Genomes to Drugs*, Germany, Wiley-VCH, 2002.

M. Stanke and S. Waack, Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies, *Nucleic Acids Res.*, **31**, 5654-5666 (2003).

StatSoft, Inc., *Electronic Statistics Textbook*, http://www.statsoft.com/textbook/stathome.html, 2008.

R.L. Tatusov, N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E.V. Koonin, D.M. Krylov, R. Mazumder, S.L. Mekhedov, A.N. Nikolskaya, B.S. Rao, S. Smimov, A>V. Sverdlov, S. Vasudevan, Y.I. Wolf , J.J. Yin and D.A. Natale, The COG database: an updated version includes eukaryotes, *BMC Bioinformatics*, **11**, 41 (2003).

J. Taylor, Clues to function in gene deserts, *Trends Biotechnol*, **23**, 269-271 (2005).

A.E. Tenney, R.H. Brown, C. Vaske, J.K. Lodge, T.L. Doering and M.R. Brent, Gene prediction and verification in a compact genome with numerous small introns, *Genome Res*., **14**, 2330-2335 (2004).

A.B. Topchy, A.K. Jain and W. Punch, A mixture model for clustering ensembles. *Proceedings of the SIAM Conference on Data Mining*, 379-390 (2004).

O.G. Troyanskaya, K. Dolinski, A.B. Owen, R. B. Altman and D. Borstein, (2003) A Bayesian framework for combining heterogenous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *PNAS*, 100, 8348-8353 (2003).

B.M. Tyler, S. Tripathy, X. Zhang, P. Dehal, R.H. Jiang R.H, A. Aerts, F.D. Arrendondo, L. Baxter, D. Bensasson, J.L. Beynon, J. Chapman, C.M. Damasceno, A.E. Dorrance, D. Dou, A.W. Dickerman, I.L. Dubchak, M. Garbelotto, M. Gijzen, S.G. Gordon, F. Govers, N.J. Grunwald, W. Huang, K.L. Ivors, R.W. Jones, S. Kamoun, K. Krampis, K.H., Lamour, M.K. Lee, W.H. McDonald, M. Medina, H.J. Meijer, E.K. Nordberg, D.J. Maclean, M.D. Ospina-Giraldo, P.F. Morris, V. Phuntumart, N.H. Putnam, S. Rash, J.K. Rose, Y. Sakihama, A.A. Salamov, A. Savidor, C.F. Scheuring, B.W. Smith, B.W. Sobral, A. Terry, T.A. Torto-Alalibo, J. Win, Z. Xu, H. Xhang, I.V. Grigoriev, D.S. Rokhsar and J.L. Boore, *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis, *Science*, **313**, 1261-1266 (2006).

UCLA Department of Statistics, *EBook*, http://www.wiki.stat.ucla.edu/socr/index.php/EBook, 2008.

A. Vanden Wymelenberg, G. Sabat, D. Martinez, A.S. Rajangam, T.T. Teeri, J. Gaskell, P.J. Kersten and D. Cullen, The *Phanerochaete chrysosporium* secretome: database predictions and initial mass spectrometry peptide identifications in cellulose-grown media, *J. Biotechnol*., **118**, 17-34 (2005).

V. Vapnik, *The Nature of Statistical Learning Theory*,Springer, New York, 1995.

M.P. Washburn, D. Wolters and J.R. Yates III, Large-scale analysis of the yeast proteome by multidimensional protein identification technology, *Nat. Biotechnol.*, **19**, 242-247 (2001).

K.M. Waters, J.G. Pounds and B.D. Thrall, Data merging for integrated microarray and proteomic analysis. *Brief. Funct. Genomics Proteomic.*, **5**, 261-272 (2006).

M. Watson, ProGenExpress: Visualization of quantitative data on prokaryotic genomes. *BMC Bioinformatics*, **6**, 98 (2005).

B.M. Webb-Robertson, E.S. Peterson, M. Singhal, K.R. Klicker, C.S. Oehmen, J.N. Adkins and S.L. Havre, PQuad – a visual analysis platform for proteomics data exploration of microbial organisms. *Bioinformatics*, **23**, 1705-1707 (2007).

C. Wei, P. Lamesch, M. Arumugam, J. Rosenberg, P. Hu, M. Vidal and M.R. Brent, Closing in on the *C. elegans* ORFeome by cloning TWINSCAN predictions, *Genome Res.*, **15**, 577-582 (2005).

D.L. Wheeler, C. Chappey, A.E. Lash, D.D. Leipe, T.L. Madden, G.D. Schuler, T.A. Tatusova and B.A. Rapp, Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **28**, 10-14 (2000).

Wikipedia, http://en.wikipedia.org/wiki/Main_Page, 2008.

Wolfram, Inc., *Mathworld Probability and Statistics*, http://mathworld.wolfram.com/topics/ProbabilityandStatistics.html, 2008.

I. Xenarios, D.W. Rice, L. Salwinski, M.K. Baron, E.M. Marcotte and D. Eisenberg, DIP: The Database of Interacting Proteins. *Nucleic Acids Res.*, **28**, 289-91 (2000).

Y. Xu, R.J. Mural and E.C. Uberbacher, Inferring gene structures in genomic sequences using pattern recognition and expressed sequence tags, *5th International Conference on Intelligent Systems for Molecular Biology*, 344-353 (1997).

H. Yao, L. Guo, Y. Fu, L.A. Borsuk, T.J. Wen, D.S. Skibbe, X. Cui, B.E. Scheffler, J. Cao, S.J. Emrich, D.A. Ashlock and P.S.Schnable, Evaluation of five *ab initio* gene prediction programs for the discovery of maize genes, *Plant Mol. Biol.*, **57**, 445-60 (2005).

E.M. Zdobnov and R. Apweiler, InterProScan-an intergration platform for the signature-recognition methods in InterPro, *Bioinformatics*, **17**, 847-848 (2001).

Y. Zeng, J. Tang, Garcia-Frias and G.R. Gao, An adaptive meta-clustering approach: combining the information from different clustering results. *Proceedings of Bioinformatics Conference*, 276-287 (2002).

K. Zhou, E.A. Panisko, J.K. Magnuson, S.E. Baker, I. Grigoriev. Proteomics for validation of automated gene model predictions. In “Mass Spectrometry of Proteins and Peptides” Eds: M Lipton, L Paša-Tolić. Humana Press. *In press.*